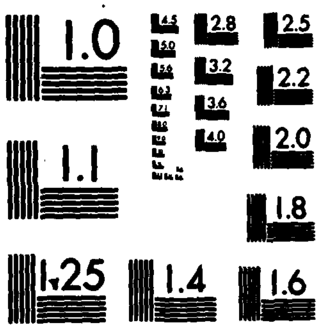
AD-A081 523 STANFORD UNIV CALIF INST FOR MATHEMATICAL STUDIES I--ETC F/G 12/1
ON PARTITIONING OF A SAMPLE WITH BINARY-TYPE QUESTIONS IN LIEU --ETC(U)
SEP 79 K J ARROW, L PESOTCHINSKY, M SOBEL N00014-79-C-0685
UNCLASSIFIED TR-295 NL

1 of 1

END
DATE
FILMED
4-80
DDC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF [illegible]

1

6

ON PARTITIONING OF A SAMPLE WITH BINARY-TYPE QUESTIONS IN LIEU OF COLLECTING OBSERVATIONS

by

10

Kenneth J. Arrow, Leon Pesotchinsky and Milton Sobel

14 TR-

9

Technical Report No. 295

September 1979

11

12 40

A REPORT OF THE
CENTER FOR RESEARCH ON ORGANIZATIONAL EFFICIENCY
STANFORD UNIVERSITY

15

Contract ONR-N00014-79-C-0685, United States Office of Naval Research

✓NSF-MSC78-01881

THE ECONOMICS SERIES
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
Fourth Floor, Encina Hall
Stanford University
Stanford, California
94305

179550

ABSTRACT

The problem is to search for the t largest observations in a random sample of size n by asking binary type questions of the people (or items) in the sample without collecting any exact data whatever. The unordered and ordered cases are both considered; in the latter case the complete ranking is of special interest. Two different criteria of optimality are considered: (1) to minimize the expected number of questions required and (2) to maximize the probability of terminating the search in at most r questions for specified r. Optimal procedures are found and compared and in some sense the solutions for these two criteria are close to each other. The analysis is nonparametric in the sense that it holds for any underlying sampling distribution but the actual optimal procedures depend on the specified distribution.

KEY WORDS

optimal sequential search

ON PARTITIONING A SAMPLE WITH BINARY-TYPE QUESTIONS IN LIEU OF COLLECTING OBSERVATIONS*

by

Kenneth J. Arrow,** Leon Pesotchinsky,***
and Milton Sobel***

1. Introduction

This problem originated in some research on the optimal design of organizations, though it clearly has many other applications. Consider the simplest problem of resource allocation, in which there is one input to be allocated among many possible users. All users produce the same one product, and each is characterized by output-input ratio independent of the scale of operations. Optimal resource allocation would require allocating the entire input to the user with the highest output-input ratio.

Suppose there are a large number of users. In the first instance, each user knows his or her own ratio only, while the center (the agent performing the allocation) does not. The center must acquire the information by asking questions of the users. However, in the spirit of information theory, the more exact the required answer the more costly its transfer is. We can reduce the problem to that of asking

*This work was supported in part by Office of Naval Research Grant ONR-N0014-79-C-0685 and in part by National Science Foundation Grant MSC78-01881. This paper started with a series of discussions between one of the present authors, Professor Kenneth Arrow of Stanford University and Professor Leonid Hurwicz of the University of Minnesota and the original applications considered were in the field of economics. Thanks are due to Professor Morris Newman of UCSB for suggesting the method of proof of Theorem 3.1. We also wish to thank Ms. Barbara Havens for carrying out some of the computations.

**Department of Economics, Stanford University

***Department of Mathematics, University of California at Santa Barbara

dichotomous questions. Since the center does not know the individual values of the output-input ratio, it may treat them as members of a random sample from a distribution. In this paper, we assume the distribution known.

There are many other situations in which the choice of the largest element from a sample may reasonably be made by binary-type questions. For instance the data to be collected has a confidential or semi-confidential nature and people may be reluctant to furnish information about their age or salary or even about the amount of money presently in their wallet. On the other hand, people may be willing to state that the quantity X in question (i.e., X equals their age) is greater than 30 and, if necessary, later tell you that it is under 45, etc. The problem (or goal) is to continue such questions with the n people (or a subset of them) in order to find the one whose X-characteristic is the most extreme in a given direction (say, the largest). A more general goal would be to find the t largest (smallest) of n with or without respect to order. For the latter goal, the case $t = n - 1$ (or n) would correspond to a complete ordering of the people in the sample and this is an important special case. Clearly, the case $t = 1$ reduces to the former goal.

The criterion to be used has to be specified exactly in order to either find the optimal procedure or decide whether a given procedure is optimal. The main criterion of interest in this paper is the expected number of questions that has to be asked. We are also interested in criteria such as "maximizing the probability of terminating in r steps." For most of our goals, the latter criterion with $r = 1$ and the main criterion (expection) give results that are in "close proximity" from the point of view of applications.

Several things should be noted:

(1) We are not allowing paired comparisons here; we do compare all x's with a single specified constant and call this one question.

(2) We assume in our illustrations that the X-characteristics of the n people, $x_1, x_2, \ldots x_n$, are independent and identically distributed (iid) (or at least exchangeable) with cdf $F(x)$, which is known to us (the case of unknown F will be considered by the authors in a separate publication).

(3) We assume that the observations (x's) are continuous so that with probability one we can assert that no two are exactly equal. We recognize that this may not be strictly true in the applications noted above and that practical modifications will be necessary to handle ties (e.g., two people may both be 45 years old and the data available to us does not give ages finer than to the nearest year). However, the theoretical analysis will not take this into account; it simply uses the fact that with probability one under very weak restrictions (independence being more than sufficient) no two x's will be exactly equal.

Remark: Moreover, there will usually be a practical lower bound to the fineness of the data, say ε, that encourages ties for large sample sizes. If we expect ties in the sample we modify our procedure by not allowing in our questions (which are of the form: "Is your X larger than c") two constants within ε of each other. Then it is easy to show that the results we have below on expectation are upper bounds for this new modified procedure, even if the probability of ties is not small.

Our solutions (for the case of known cdf $F(x)$) are strongly dependent on the given cdf $F(x)$ (i.e., when the true cdf $F(x)$ is completely specified). However, the solutions are nonparametric in the sense that the instructions and tables needed to carry out the procedures are the same regardless of the particular assumed $F(x)$. Thus our tables would specify a value of p and the procedure (at the first step) might be to solve $F(c) = p$. Another equivalent way of stating this is that the problem has been reduced to that of the uniform $(0,1)$ distribution.

The results obtained are quite striking. Thus in the basic illustration $(t = 1)$ the minimal expected number of questions required is less than 2 1/2, namely 2.42778. The result above holds for any starting sample size n and for any known cdf $F(x)$. The procedure that maximizes the probability of terminating on the very next step (the second criterion with $r = 1$) has a result not far removed, namely the corresponding expected number of questions required is 2.44144. The optimal procedure in the latter sense is simpler because it does not require the use of any table of optimal p(or c)-values.

Of course, there could be some other ways of asking group questions, e.g., rather than asking a binary type question leading to a yes or no answer (such as "raise your hand if your X is larger than c and do not raise otherwise"), we could allow questions with 3 possible answers (such as "raise your right hand (or red flag) if $X > c_2$, your left hand (or blue flag) if $X \leqq c_1$, and no hand (flag) otherwise"). Then with $c_1 < c_2$ we can partition the sample with one question into at most 3 disjoint sets: $X \leqq c_1$, $c_1 < X \leqq c_2$, $X > c_2$. In the same way questions with $k_0(> 3)$ possible answers may be allowed and, of course, we should make every attempt to use such questions if we wish to attain an optimal solution. (The reason for this is that for $k_2 > k_1 \geqq 2$ an optimal procedure with "k_2-way" questions generally gives better results than an optimal procedure with "k_1-way" questions.) In the illustrations below only binary type questions are allowed; however, the same approach could be implemented in the cases of more complicated "sampling procedures" (we refer to the type of question allowed as a part of our "sampling procedure").

We regard our problem as the partial or complete ordering of a sample without the necessity of knowing any particular values of the observations in the sample.

In Section 2 we consider the problem of selecting without order the t largest of n observations in a random sample; the same problem with ordering is discussed at the end of Section 3. The main part of Section 3 deals with the problem of a complete ordering of the sample.

2. Selecting Without Order the t Largest

2.1 Preliminaries

Consider the problem of selecting without order the t largest observations in a sample (say, of people) of size n when the above type of sampling is available to us, i.e., we can ask any subset of the n people to each raise his hand if (and only if) his $X > c$, where c is at our disposal to select. We terminate when (and only when) we definitely have the t largest separated from all the others. (The modifications required in the case of ties will be evident in the light of a remark made in Section 1 above.)

It should be understood that if we obtain a subset of size k which is less than t as a result of the first question then we continue looking for $t - k$ from the batch of size $n - k$; if $k > t$ then we continue looking for t from the reduced batch of size k.

Let $\pi_{i,j}$ denote the probability that j people out of i will respond affirmatively to a single question. In fact these $\pi_{i,j}$-values can in the most general set-up depend on the entire history of the procedure. In other words, if $\{\omega\}$ denotes the space of trajectories of a random process associated with our procedure, then after w questions have been asked (or at a moment w) $\pi_{i,j}(\omega) = \pi_{i,j}(\omega_s, s \leqq w)$ for $\omega_w = i$. However, using the assumption that $\pi_{i,j} > 0$ for $j \leqq i$, it can be shown in a manner completely analogous to that given in a book by Ross [1970, Ch. 6] that if the X's are independent (or at least exchangeable), then the optimal solution for our principal criterion (minimum expected number of questions) is obtained by a stationary Markov decision procedure, i.e., with transition probabilities $\pi_{i,j}$ which do not depend on ω.

As a result of the above we can use the Markovian property to write for $n \geqq 2$ the basic equation for our procedure

$$(2.1) \qquad P_{n,t}(r) = \sum_{j=t+1}^{n} \pi_{n,j} P_{j,t}(r-1) + \sum_{j=0}^{t-1} \pi_{n,j} P_{n-j,t-j}(r-1) \quad ,$$

where $P_{n,t}(r)$ denotes the probability of terminating in at most r steps under a scheme described above if we start with n and our goal is to find the top t unordered. The boundary condition is simply that for all t we should have $P_{t,t}(r)$ equal to zero for $r \geqq 1$ and equal to one for $r = 0$. If we multiply both sides of (2.1) by $r - 1$ and sum from $r = 1$ to ∞ (letting $\mu_n^{(t)}$ denote the expected number of steps or questions required for our procedure with n and t defined in the present goal), then we easily obtain

(2.2) $$\mu_n^{(t)} = \frac{1 + \sum_{j=t+1}^{n-1} \pi_{n,j}\mu_j^{(t)} + \sum_{j=1}^{t-1} \pi_{n,j}\mu_{n-j}^{(t-j)}}{1 - \pi_{n,0} - \pi_{n,n}} \quad .$$

A sufficient condition for the expected time of absorption (which is equivalent to the expected number of questions needed in our problems) to be finite and hence for the both sides of (2.2) to be finite is that the $\pi_{i,j}$ be bounded away from zero; more precisely that $\pi_{i,j} > \delta$ for all i and j $(j \leqq t < i \leqq n)$ for some $\delta > 0$. (This sufficient condition can be shown to hold for our optimal procedure.) An alternative way to show that $\mu_n^{(t)}$ are all finite for the case of the optimal procedure is to come up with some other (i.e., any) procedure for which the expected times of absoption are finite for all n. In fact, it will be shown that an optimal procedure in the sense of our second criterion with $r = 1$ yields finite values for all $\mu_n^{(t)}$.

Assuming now that the transition probabilities (i.e., the $\pi_{n,j}$) are defined by a finite number v of parameters $p_1(n),\ldots,p_v(n)$, and that $\mu_i^{(s)}$ have been found for all $s \leqq t$ and $i \leqq n - 1$, equation (2.2) can be viewed as a recurrence relation from which the $\pi_{n,j}$ and the optimal $p_1(n),\ldots,p_v(n)$ can be found via minimization of $\mu_n^{(t)}$. That defines an optimal procedure in the sense of our basic criterion (minimizing the expectation). For our second criterion with $r \geq 2$ (i.e., maximization of $P_{n,t}(r)$ for a specific r), we can treat equation (2.1) in an analogous manner assuming that the $P_{i,s}(\tau)$ have been found for all $i \leq n$, $s \leq t$ and $\tau \leq r - 1$. For $r = 1$ the solution in the latter case is obtained simply by maximization of the single coefficient $\pi_{n,t}$.

In the important special case where $x_1,\ldots,x_n$ are obtained from continuous iid random variables, the transition probabilities $\pi_{n,j}$ are binomial, namely $\pi_{n,j} = \binom{n}{j}p_n^j(1 - p_n)^{n-j}$, where $p_n = 1 - F(c_n)$ and c_n is a value which defines the initial question for the sample size n. Therefore, in this setting the optimal solution is given by a sequence of values $p_n^{(\tau)}$, $n \geq \tau + 1$, $\tau \leq t$. Naturally, since the search for the top t unordered is equivalent to that for the bottom $n - t$ unordered, we can assume that $n \geq 2t$. For the second criterion with $r - 1$ we obtain $\tilde{p}_n$ values maximizing $\pi_{n,t} = \binom{n}{t}p_n^t(1 - p_n)^{n-t}$; it is easily shown that the sequence $\tilde{p}_n = t/n$ is optimal in the sense of this criterion. We can see now that $\pi_{n,t} \downarrow t^t e^{-t}/t!$ as $n \to \infty$ and hence in the associated Markov chain the absorption time $\tilde{\mu}_n^{(t)} \leq (\underline{\lim}\, \pi_{n,t})^{-1} = t!e^t/t^t$ and hence $\tilde{\mu}_\infty^{(t)} = \overline{\lim}\, \tilde{\mu}_n^{(t)}$ satisfies the same inequality. Thus for each t the quantity $t!(e/t)^t$ is an upper bound on the optimal $\mu_n^{(t)}$ for any n; this upper bound holds both for optimality of the second criterion with $r = 1$ as well as for that of the basic criterion.

Now we can use (2.2) to derive some results for the optimal procedure with respect to the basic criterion.

2.2 The Optimal Procedure for the Basic Criterion

Theorem 2.1: Let $\{p_n^{(t)}\}$ be optimal in the sense of our basic criterion and $\mu_n^{(t)}$ denotes the corresponding expectations. Then

(1) $\mu_n^{(t)}$ increases with n (for any fixed t) and

$\mu_n^{(t)} \leqq \mu_\infty^{(t)} < \tilde{\mu}_\infty^{(t)} < \sqrt{2\pi t}\, e^{1/12t}$;

(2) $t/n \leqq p_n^{(t)} < (t+1)/(n+1)$ for $t < (n+1)/2$ and

$\lim_{n\to\infty} np_n^{(t)} = \theta_t$ exists $(t < \theta_t < t+1)$.

Proof: For simplicity let us first take $t = 1$ and write p instead of p_n (or $p_n^{(t)}$), q instead of $1 - p$ and μ_n instead of $\mu_n^{(t)}$. Then (2.2) can be written as

$$\mu_n(p) = \frac{1 + \sum_{j=2}^{n-1} \binom{n}{j} p^j q^{n-j} \mu_j}{1 - p^n - q^n} \tag{2.3}$$

which defines an analytic function of p on $(0,1)$, so that for the optimal p (i.e., $p = p_n^{(1)}$), we have $(d\mu_n(p))/dp = 0$ and it gives us

$$(2.4)\qquad \sum_{j=2}^{n-1}\binom{n}{j}p^{j-1}q^{n-1-j}(j - np)\mu_j$$

$$= n(q^{n-1} - p^{n-1})\left(1 + \sum_{j=2}^{n-1}\binom{n}{j}p^j q^{n-j}\mu_j\right)(1 - p^n - q^n)^{-1}$$

$$= n(q^{n-1} - p^{n-1})\mu_n(p) = n(q^{n-1} - p^{n-1})\mu_n \quad .$$

Using the identity $j - np = nq - (n - j)$ in the left-hand side of (2.4) we then rewrite (2.4) in the form

$$(2.5)\qquad \frac{n}{p}\left\{\sum_{j=2}^{n-1}\binom{n}{j}p^j q^{n-j}\mu_j - \sum_{j=2}^{n-2}\binom{n-1}{j}p^j q^{n-1-j}\mu_j - p^{n-1}\mu_{n-1}\right\}$$

$$= n(q^{n-1} - p^{n-1})\mu_n \quad .$$

For any p and the optimal μ_{n-1} from (2.3) it is clear that

$$(2.6)\qquad \mu_{n-1} \leq \left(1 + \sum_{j=2}^{n-2}\binom{n-1}{j}p^j q^{n-1-j}\mu_j\right)(1 - p^{n-1} - q^{n-1})^{-1} \quad .$$

Hence we obtain from (2.3), (2.5) and (2.6)

$$(2.7)\qquad \mu_n(pq^{n-1} - p^n) \leq \mu_n(1 - p^n - q^n) - \mu_{n-1}(1 - q^{n-1}),$$

which yields

$$\mu_{n-1} \leqq \mu_n \quad . \tag{2.8}$$

Now, in the first line of (2.4) we expand $n - np$ as two separate terms and use (2.3) on the second sum. Equating this to the third term in (2.4) we obtain, for any p_r that satisfies $(d\mu_n(p))/dp = 0$,

$$[\mu_n(p_r) - \mu_n](1 - p_r^{n-1}) = 1 - \mu_n + \sum_{j=2}^{n} \binom{n-1}{j-1} p_r^{j-1} q_r^{n-j} \mu_j \quad . \tag{2.9}$$

By virtue of (2.8) the right side of (2.9) is an increasing function of p_r and hence by (2.9) $\mu_n(p_r)$ is also, i.e., if there are two solutions p_1, p_2 with $p_1 < p_2$ then $\mu_n(p_1) \leqq \mu_n(p_2)$. Since $\mu_n(p)$ is an analytic function in $(0,1)$ and tends to infinity as $p \to 1$ from the left and also as $p \to 0$ from the right, there cannot be two such p-values that are both local minima. Hence the minimum must be unique and there is no analytical maxima in $(0,1)$.

The next task is to locate this unique minimum. From the first and third terms in (2.4)

$$\frac{d\mu_n(p)}{dp} = \frac{\sum_{j=2}^{n-1} \binom{n}{j} (j - np) p^{j-1} q^{n-1-j} \mu_j - n\mu_n(p)(q^{n-1} - p^{n-1})}{1 - p^n - q^n} \quad . \tag{2.10}$$

We wish to show that for every n the derivative is negative at $p = 1/n$ and positive at $p = 2/(n + 1)$. Firstly, for $p = 1/n$ we take into account that $\mu_n(p) \geqq \mu_n \geqq \mu_j > \mu_2 = 2$ for $j > 2$ (the fact that $\mu_2 = 2$ can easily be derived either directly or from the associated Markov chain). We first separate the term with $j = 2$ in the sum in the numerator of (2.10) and then in the remaining terms expand $j - np = j - 1$ as two separate terms, thus obtaining for $n > 2$ that the numerator of (2.10) for $p = 1/n$ does not exceed

$$(n - 1)q^{n-3} = n\mu_n \left\{ q^{-1}\sum_{j=3}^{n-1} \binom{n-1}{j-1} p^{j-1}q^{n-j} - q^{-1}\sum_{j=3}^{n-1} \binom{n}{j} p^{j}q^{n-j} - q^{n-1} + p^{n-1} \right\}$$

$$= (n - 1)q^{n-3}\{1 - \mu_n/2\} < 0 \quad ,$$

so that

$$\left. \frac{d\mu_n(p)}{dp} \right|_{p = 1/n} < 0 \quad .$$

In the same way that (2.9) has been derived from (2.4), we can use (2.10) and some algebra to show that

$$(2.11) \quad \frac{d\mu_n(p)}{dp}$$

$$= n\{p(1 - p^n - q^n)\}^{-1}\{[\mu_n(p) - \mu_{n-1}(p)]$$

$$\cdot (1 - q^{n-1}) + p^{n-1}[\mu_{n-1}(p) - \mu_{n-1}]\} \ .$$

We can see now from (2.11) that if $\mu_n(p) \geqq \mu_{n-1}(p)$, then $(d\mu_n(p))/dp > 0$ and with the continuity of $\mu_n(p)$ in p this implies that $p_n^{(1)} < p$ if $\mu_n(p) - \mu_{n-1}(p) \geqq 0$. It now remains only to show that the latter inequality holds for $p = 2/(n + 1)$. But using (2.2) and the relation $\pi_{n-1,k} = \pi_{n,k}(n - k)/nq$ we can show that

$$(2.12) \qquad (1 - p^n - q^n)(1 - p^{n-1} - q^{n-1})[\mu_n(p) - \mu_{n-1}(p)]$$
$$= -pq^{n-1} - qp^{n-1} + \sum_{k=2}^{n-2} \pi_{nk}\mu_k[1 - p^{n-1} - q^{n-1} - \frac{(1 - p^n - q^n)(n - k)}{nq}]$$
$$+ np^{n-1}q\mu_{n-1}(1 - p^{n-1} - q^{n-1})$$

and the terms in the sum in the right-hand side of (2.12) are non-negative if $k \geqq 3$ and $p = 2/(n + 1)$. Since for $k \geqq 3$ we have $\mu_k > \mu_2 = 2$, the right-hand side in (2.12) is not smaller than

$$-pq^{n-1} - qp^{n-1} + 2\{(1 - p^n - q^n - npq^{n-1})(1 - p^{n-1} - q^{n-1})$$
$$- (1 - p^{n-1} - q^{n-1} - (n - 1)pq^{n-2})(1 - p^n - q^n)\}$$
$$= -pq^{n-1} - qp^{n-1} + 2pq^{n-2}(np - 1)$$

and with $p = 2/(n + 1)$ the latter expression equals $pq^{n-1} - qp^{n-1} > 0$. This implies that

$$(2.13) \qquad \frac{1}{n} \leqq p_n^{(1)} < \frac{2}{n + 1}$$

and the equality holds only for $n = 2$. The inequality (2.13) enables us to take limits in (2.3) in accordance with two subsequences converging respectively to $\underline{\lim}\, np_n^{(1)}$ and $\overline{\lim}\, np_n^{(1)}$ (using the Poisson approximation to the binomial distribution). The Poisson approximation and the monotonicity of $\mu_n^{(1)}$ imply that $\underline{\theta}_1 = \underline{\lim}\, np_n^{(1)} = \overline{\lim}\, np_n^{(1)} = \overline{\theta}_1 = \theta_1$ (say). A new equation can now be written instead of (2.3), namely

$$(2.14)\quad \mu_\infty^{(1)} = \left\{1 + e^{-\theta_1} \sum_{i=2}^{\infty} \frac{(\theta_1)^i}{i!} \mu_i^{(1)}\right\} (1 - e^{-\theta_1})^{-1}$$

which is more convenient for the numerical search for θ_1. This completes the proof for $t = 1$.

For $t > 1$ the proof of Theorem 2.1 follows essentially in the same manner as above and hence we will outline the result below.

Firstly, two identities can be derived from (2.2) in the same manner as (2.4) and (2.10) are derived from (2.3), namely

$$(2.15)\quad \frac{d\mu_n^{(t)}(p)}{dp} = n\{p(1 - p^n - q^n)\}^{-1}\{[\mu_n^{(t)}(p) - \mu_{n-1}^{(t)}(p)](1 - q^{n-1})$$
$$+ p^{n-1}[\mu_{n-1}^{(t)}(p) - \mu_{n-1}^{(t)}] - \sum_{k=1}^{t-1} \pi_{n-1,k}(\mu_{n-k}^{(t-k)} - \mu_{n-k-1}^{(t-k)})\}$$

and

$$(2.16)\quad \frac{d\mu_n^{(t)}(p)}{dp} = n\{q(1 - q^n - p^n)\}^{-1}\left\{\sum_{k=t}^{n-2}\pi_{n-1,k}(\mu_{k+1}^{(t)} - \mu_k^{(t-1)}) - q^{n-1}[\mu_{n-1}^{(t-1)}(p) - \mu_{n-1}^{(t-1)}] - [\mu_n^{(t)}(p) - \mu_{n-1}^{(t-1)}(p)] \cdot (1 - p^{n-1})\right\} .$$

Since in the point $p_n^{(t)}$ we have $(d\mu_n^{(t)}(p))/dp = 0$, we can see from (2.15) and the fact that $\mu_{n-1}^{(t)} \leqq \mu_{n-1}^{(t)}(p)$ for all p that

$$(2.17)\quad \sum_{k=1}^{t-1}\pi_{n-1,k}(\mu_{n-k}^{(t-k)} - \mu_{n-k-1}^{(t-k)})$$
$$= [\mu_n^{(t)} - \mu_{n-1}^{(t)}(p)](1 - q^{n-1}) + p^{n-1}[\mu_{n-1}^{(t)}(p) - \mu_{n-1}^{(t)}]$$
$$= [\mu_n^{(t)} - \mu_{n-1}^{(t)}(p)](1 - p^{n-1} - q^{n-1}) + p^{n-1}[\mu_n^{(t)} - \mu_{n-1}^{(t)}]$$
$$\leqq [\mu_n^{(t)} - \mu_{n-1}^{(t)}](1 - q^{n-1})$$

and under the inductional assumption that $\mu_r^{(v)} > \mu_{r-1}^{(v)}$ for all $r < n$ and $v < t$ (2.17) implies that $\mu_n^{(t)} > \mu_{n-1}^{(t)}$. Then we can prove that the differences $\mu_{k+j}^{(j)} - \mu_{k+j-1}^{(j)}$ are increasing with j which in turn implies that $\mu_{t+\alpha}^{(t)} - \mu_{t-1+\alpha}^{(t-1)}$ increases with $\alpha (\alpha \geqq 1)$ because $\mu_{t+\alpha}^{(t)} - \mu_{t-1+\alpha}^{(t-1)} = \mu_{t+\alpha}^{(\alpha)} - \mu_{t-1+\alpha}^{(\alpha)}$.

For any point p in which $(d\mu_n^{(t)}(p))/dp = 0$ we can write now from (2.15) that

$$(2.18) \qquad \mu_n^{(t)}(p) - \mu_{n-1}^{(t)} = (\mu_{n-1}^{(t)}(p) - \mu_{n-1}^{(t)})\frac{1 - p^{n-1} - q^{n-1}}{1 - q^{n-1}} + \frac{\sum_{k=1}^{t-1} \pi_{n-1,k}(\mu_{n-k}^{(t-k)} - \mu_{n-1-k}^{(t-1)})}{1 - q^{n-1}}$$

and from (2.16) that

$$(2.19) \qquad \mu_n^{(t)}(p) - \mu_n^{(t)} = -\frac{\mu_n^{(t)} - \mu_{n-1}^{(t-1)}}{1 - p^{n-1}} + (\mu_{n-1}^{(t-1)}(p) - \mu_{n-1}^{(t-1)})\frac{1 - p^{n-1} - q^{n-1}}{1 - p^{n-1}} + \frac{\sum_{k=t}^{n-1} \pi_{n-1,k}(\mu_{k+1}^{(t)} - \mu_k^{(t - 1)})}{1 - p^{n-1}} .$$

Suppose now that a function $\mu_r^{(v)}(p)$ decreases in the interval $(0,p_r^{(v)})$ and increases in the interval $(p_r^{(v)},1)$ and that $v/r \leq p_r^{(v)} < (v + 1)/(r + 1)$ for all $v \leq t$, $r \leq n-1$ and $v < (1 + r)/2$.

Then the right-hand side in (2.18) decreases in the interval $(0,p_{n-1}^{(t)})$ (because $\mu_{n-1}^{(t)}(p)$ decreases by virtue of the inductional assumption and $\mu_{n-k}^{(t-k)} - \mu_{n-1-k}^{(t-k)}$ decreases with k due to the statement above after (2.17)). At the same time, the right-hand side in (2.19) increases in the interval $(p_{n-1}^{(t-1)},1)$ and with the help of the arguments following (2.9) we can see that there cannot be two such p-values to the left of $p_{n-1}^{(t)}$ (to the right of $p_{n-1}^{(t-1)}$) that are both local minima. But by virtue of the inductional assumption

$$p_{n-1}^{(t-1)} < \frac{t}{n} < \frac{t}{n-1} < p_{n-1}^{(t)} \quad ,$$

which implies that if there is a local minimum in the interval $[p_{n-1}^{(t-1)}, p_{n-1}^{(t)}]$, then this minimum is unique for the function $\mu_n^{(t)}(p)$ in $(0,1)$. From (2.15) we obtain now that

$$\left.\frac{d\mu_n^{(t)}(p)}{dp}\right|_{p = p_{n-1}^{(t)}}$$

$$= n\{p(1 - p^n - q^n)\}^{-1}\Big\{(\mu_n^{(t)}(p) - \mu_{n-1}^{(t)})(1 - q^{n-1})$$

$$- \sum_{k=1}^{t-1} \pi_{n-1,k} (\mu_{n-k}^{(t-k)} - \mu_{n-k-1}^{(t-k)})\Big\}$$

$$\geqq n\{p(1-p^n - q^n)\}^{-1}(\mu_n^{(t)} - \mu_{n-1}^{(t)})\left(1 - q^{n-1} - \sum_{k=1}^{t-1} \pi_{n-1,k}\right) > 0 \quad ,$$

so that $\mu_n^{(t)}(p)$ increases in the interval $(p_{n-1}^{(t)},1)$. In a similar manner we can show that

$$\left.\frac{d\mu_n^{(t)}(p)}{dp}\right|_{p=p_{n-1}^{(t-1)}} < 0$$

and hence there is a unique minimum of $\mu_n^{(t)}(p)$, and that it lies in the interval $(p_{n-1}^{(t-1)},p_{n-1}^{(t)})$. then for $t \geqq 2$ we can show that $(d\mu_n^{(t)}(p))/dp$ is not positive in the point $\underline{\rho}$ of the intersection of $\mu_n^{(t)}(p)$ and $\mu_{n-1}^{t}(p)$ and $(d\mu_n^{(t)}(p))/dp$ is positive in the point $\overline{\rho}$ of intersection of $\mu_n^{(t)}(p)$ and $\mu_{n-1}^{(t-1)}(p)$, so that $\underline{\rho} < p_n^{(t)} < \overline{\rho}$. The final step of the proof includes the direct verification of the fact that $\underline{\rho} \geqq t/n$ and $\overline{\rho} \leqq (t+1)/(n+1)$ (which we omit) and hence

$$(2.20) \qquad t/n \leqq p_n^{(t)} < (t+1)/(n+1) \quad .$$

It is easy to show that the equality in (2.20) holds only if $n = 2t$. Indeed, if $n = 2t$ we obtain from (2.2) (taking into account that $\mu_{2t-k}^{(t-k)} = \mu_{2t-k}^{(t)}$) that

(2.21) $$\mu_{2t}^{(t)}(p) = \frac{1 + \sum_{k=1}^{t-1} \mu_{t+k}^{(t)}(\pi_{2t,t+k} + \pi_{2t,t-k})}{1 - p^{2t} - q^{2t}}$$

and the monotonicity of $\mu_{t+k}^{(t)}$ in k implies that $p = 1/2$ yields the minimum for $\mu_{2t}^{(t)}(p)$.

The proof of the existence of

$$\lim np_n^{(t)} = \theta_t \quad , \quad t \leqq \theta_t < t + 1 \quad ,$$

is similar to that in the case $t = 1$. ■

2.3 Approximation to the Optimal Procedure.

The above results enable us to prove that the proximity of t/n and $p_n^{(t)}$ implies the same for $\mu_n^{(t)}$ and $\tilde{\mu}_n^{(t)}$; the latter refers to the second criterion discussed in Section 2.1 above.

Theorem 2.2: There exists an $\varepsilon > 0$ such that for all $n \leqq \infty$ and $t < (n + 1)/2$

$$\tilde{\mu}_n^{(t)} - \mu_n^{(t)} \leqq \varepsilon .$$

Proof: We can write that

(2.22) $$\tilde{\mu}_n^{(t)} - \mu_n^{(t)} = \{\tilde{\mu}_n^{(t)} - \bar{\mu}_n^{(t)}\} + \{\bar{\mu}_n^{(t)} - \mu_n^{(t)}\} ,$$

where $\bar{\mu}_n^{(t)}$ is defined by (2.2) with optimal $\mu_s^{(u)}$, $u \leqq t$, $s \leqq n - 1$, and $p = t/n$. The first term in (2.22) is an "improvement" introduced to the optimal procedure in the sense of the second criterion with $r = 1$ by substituting $\tilde{\mu}_s^{(u)}$ for the smaller values $\mu_s^{(u)}$ and the second term is an "improvement" to $\bar{\mu}_n^{(t)}$ due to the optimal choice of p.

Let us suppose now that the statement of the theorem holds for all $u \leqq t - 1$. Then we get from (2.21) (taking into account that $p_{2t}^{t} = 1/2$ and denoting $\sup\limits_{2u-1<s;u\leqq t-1} \{\tilde{\mu}_s^{(u)} - \mu_s^{(u)}\}$ by ε_{t-1}) that

$$(2.23) \qquad \tilde{\mu}_{2t}^{(t)} - \mu_{2t}^{(t)} = \tilde{\mu}_{2t}^{(t)} - \bar{\mu}_{2t}^{(t)}$$

$$= \frac{\sum\limits_{k=1}^{t-1} (\tilde{\mu}_{t+k}^{(k)} - \mu_{t+k}^{(k)})(\pi_{2t,t+k} + \pi_{2t,t-k})}{1 - p^{2t} - q^{2t}}$$

$$\leqq \varepsilon_{t-1} - \varepsilon_{t-1} \cdot \frac{\binom{2t}{t} 2^{-2t}}{1 - 2^{-2t+1}} .$$

The inequality above serves as the basis for the induction. If we suppose now that the statement of the theorem holds for all $u \leqq t$ and $s \leqq n - 1$ with $\varepsilon = \varepsilon_{s-1}$, then from (2.2) we can obtain that

$$(2.24) \qquad \tilde{\mu}_n^{(t)} - \bar{\mu}_n^{(t)} \leqq \varepsilon - \varepsilon \frac{\binom{n}{t}(\frac{t}{n})^t(1 - \frac{t}{n})^{n-t}}{1 - (\frac{t}{n})^n - (1 - \frac{t}{n})^n} \; .$$

In order to estimate the second term in (2.22) we can derive (2.16) that in the interval $(t/n, p_n^{(t)})$

$$(2.25) \qquad \left|\frac{d\mu_n^{(t)}(p)}{dp}\right| \leqq \frac{nC_t q^{n-1}}{1 - p^n - q^n} \; ,$$

where C_t is a constant for a fixed t and $C_t \leqq C_0 t$, where C_0 does not depend on t and $C_0 < 1$.

From (2.24) and (2.25) we can see now (using the inequality

$$\bar{\mu}_n^{(t)} - \mu_n^{(t)} \leqq \max_{\frac{t}{n} \leqq p < p_n^{(t)}} \left|\frac{d\mu_n^{(t)}(p)}{dp}\right| (p_n^{(t)} - \frac{t}{n})$$

that

$$(2.26) \qquad \tilde{\mu}_n^{(t)} - \mu_n^{(t)} \leqq \varepsilon - \frac{\varepsilon\binom{n}{t}(\frac{t}{n})^t(1 - \frac{t}{n})^{n-t} - C_0 t(1 - \frac{t}{n})^{n-t}}{1 - (\frac{t}{n})^n - (1 - \frac{t}{n})^n}$$

and

$$\tilde{\mu}_n^{(t)} - \mu_n^{(t)} \leqq \varepsilon$$

if

$$\varepsilon\binom{n}{t}\left(\frac{t}{n}\right)^{t} > C_0 t \quad ,$$

or if

$$(2.27) \qquad \varepsilon > \alpha_t = C_0\sqrt{2\pi}\, t^{3/2} e^{-t} > C_0\binom{n}{t}^{-1} n^{t} t^{-t+1} \quad .$$

Since $\alpha_t \to 0$ as $t \to \infty$ the ε can always be chosen to satisfy the statement of the theorem.

The calculations presented in the Table 1 show that for $t \leqq 7$ we have $\varepsilon_t \leqq \varepsilon_1 \leqq 0.014$ and since $\alpha_8 < 0.014$ we have $\varepsilon = \varepsilon_1 \leqq 0.014$.

From Theorem 2.2 and Table 1 we can see that from a practical point of view the optimal procedure for the second criterion (with $r = 1$) is nearly optimal in the sense of the first criterion. Clearly, the so-called t/n-procedure is more convenient since it does not require specific tables for determining the $p_n^{(t)}$'s.

Table 1

Selecting Without Order the t Largest

t	n	$p_n^{(t)}$	$\mu_n^{(t)}$	$\tilde{\mu}_n^{(t)}$	ϵ_t
1	2	.5	2.0	2.0	0
	3	.34627	2.16507	2.16667	.00160
	4	.26557	2.23783	2.24138	.00355
	10	.11111	2.35625	2.36524	.00899
	50	.02281	2.41389	2.42676	.01287
	∞	1.14852*	2.42778	2.44144	.01366
2	4	.5	2.38004	2.38095	.00091
	5	.40582	2.47956	2.48139	.00183
	6	.34173	2.53757	2.54042	.00285
	10	.20969	2.63849	2.64418	.00569
	50	.04318	2.73979	2.74998	.01019
	∞	2.17566	2.76250	2.77329	.01079
3	6	.5	2.59500	2.59706	.00206
	7	.43187	2.66373	2.66642	.00269
	8	.38017	2.70961	2.71295	.00334
	10	.30683	2.76729	2.77175	.00446
	50	.06326	2.91523	2.92435	.00912
	∞	3.18865	2.94586	2.95512	.00926
4	8	.5	2.74003	2.74307	.00304
	9	.44657	2.79158	2.79508	.00350
	10	.40350	2.82890	2.83286	.00496
	15	.27236	2.92477	2.93045	.00568
	50	.08324	3.03050	3.03913	.00863
	∞	4.19669	3.06868	3.07700	.00832
5	10	.5	2.84750	2.85131	.00381
	11	.45603	2.88829	2.89246	.00417
	12	.41919	2.91945	2.92397	.00452
	15	.33748	2.98061	2.98601	.00540
	50	.10317	3.11403	3.12240	.00837
	∞	5.20228	3.15957	3.16726	.00769
6	12	.5	2.93189	2.93634	.00445
	13	.46263	2.96538	2.97011	.00473
	14	.40348	2.99198	2.99697	.00499
	20	.30388	3.08103	3.08726	.00623
	50	.12307	3.17819	3.18641	.00822
	∞	6.20649	3.23100	3.23823	.00723
7	14	.5	3.0083	3.00579	.00496
	15	.46751	3.02909	3.03427	.00518
	20	.35294	3.11390	3.12003	.00613
	30	.23691	3.18262	3.18984	.00722
	50	.14295	3.22936	3.23751	.00815
	∞	7.20982	3.28944	3.29631	.00687

* For $n = \infty$ the entry in the $p_n^{(t)}$ column represents $\lim_{n \to \infty} np_n^{(t)}$.

3. The Complete Ordering of a Sample

Let us denote by G_n the expected number of required steps. Then in the same manner as above for $\mu_n^{(t)}$ we can write

$$(3.1) \qquad G_n(p) = \frac{1 + \sum_{r=1}^{n-1} \binom{n}{r} p^r q^{n-r} (G_r + G_{n-r})}{1 - p^n - q^n} \quad .$$

In order to agree on one definite procedure (out of many equivalent procedures) after the first question has been asked, it is understood that we shall first order the r subjects in one of the two subgroups formed and then order the remaining $n - r$ subjects. Thus the minimization of (3.1) will provide the optimal results in the sense of expectation.

We can show algebraically that for $n \leqq 5$ the optimal value of p_n is 1/2. For many values of $n \geqq 6$ it is no longer true; however, the procedure with $p_n = 1/2$ for all n serves in this case as an approximation to the optimal procedure, just like the "t/n-procedure" from Section 2 does in the case of selecting without order the t largest. Unlike the previous problem we have not obtained exact analytic results on the limiting approach of the "1/2-procedure" to the optimal procedure, but we do have considerable empirical information about this which we will describe later. What we do have is an explicit upper bound for $n \geqq 6$ for the optimal procedure which is based on the 1/2-procedure. Furthermore, the numerical results in Table 2 indicate that the difference between the G_n-results for $p = 1/2$ (denoted by $\tilde{G}_n$) and the optimal G_n-value is extremely small.

Theorem 3.1: For the 1/2-procedure we have the exact result

$$(3.2) \qquad \tilde{G}_n = \sum_{r=2}^{n} (-1)^r \binom{n}{r} (r-1)(1-2^{1-r})^{-1} = \frac{n-1}{\ell n2} + \alpha_n$$

where $0 \leqq \alpha_n \leqq 1/2$ $(n = 2,3,\ldots)$.

Proof: From (3.1) with $p = 1/2$ we have

$$(3.3) \qquad 2^n \tilde{G}_n = 2^n + 2\sum_{r=0}^{n} \binom{n}{r} \tilde{G}_r - 2(\tilde{G}_0 + n\tilde{G}_1) \quad ,$$

where we define $\tilde{G}_0 = \tilde{G}_1 = 1$ in order for (3.3) to hold for $n = 0, 1$. Multiplying through by $z^n/n!$, we let

$$y(z) = \sum_{n=0}^{\infty} \frac{\tilde{G}_n}{n!} z^n$$

and obtain

$$(3.4) \qquad y(2z) = e^{2z} + 2y(z)e^z - 2\sum_{n=0}^{\infty} \frac{n+1}{n!} z^n$$

$$= e^{2z} + 2y(z)e^z - 2(z+1)e^z \quad .$$

Table 2

Complete Ordering

n	Optimal p_n	Optimal G_n	½-procedure $\tilde{G}_n$	Asymptotic $\frac{n}{\ln 2} - 1$	Difference $\frac{n}{\ln 2} - 1 - G_n$
2	.50000	2.00000	2.00000	1.88539	-.11461
3	.50000	3.33333	3.33333	3.32808	-.00525
4	.50000	4.76190	4.76190	4.77078	.00888
5	.50000	6.20952	6.20952	6.21347	.00395
6	.53686	7.65650	7.65653	7.65617	-.00033
7	.61306	9.09874	9.10046	9.09887	.00013
8	.63462	10.54059	10.54268	10.54156	.00097
9	.64504	11.98319	11.98451	11.98425	.000106
10	.50000	13.42597	13.42660	13.42695	.00098
11	.50000	14.86811	14.86909	14.86965	.00154
12	.50000	16.31053	16.31188	16.31234	.00181
13	.50000	17.75312	17.75481	17.75503	.00191
14	.50000	19.19579	19.19775	19.19773	.00194
15	.50000	20.63847	20.64062	20.64042	.00195
16	.58992	22.08108	22.08341	22.08312	.00204
17	.62204	23.52360	23.52610	23.52582	.00222
18	.63860	24.96611	24.96873	24.96851	.00240
19	.64838	26.40864	26.41133	26.41120	.00256
20	.65530	27.85120	27.85391	27.85390	.00270
21	.66184	29.29377	29.29651	29.29659	.00282
22	.67052	30.73636	30.73914	30.73930	.00294
23	.68381	32.17894	32.18179	32.18199	.00305
24	.50000	33.62148	33.62447	33.62468	.00320
25	.50000	35.06402	35.06719	35.06738	.00336
26	.50000	36.50657	36.50992	36.51007	.00350
27	.50000	37.94912	37.95266	37.95276	.00364
28	.50000	39.39169	39.39542	39.39546	.00377
29	.50000	40.83425	40.83817	40.83815	.00390
30	.50000	42.27682	42.28092	42.28085	.00403
31	.50000	43.71939	43.72366	43.72355	.00416
32	.55179	45.16196	45.16639	45.16625	.00429
33	.57485	46.60452	46.60911	46.60894	.00442
34	.58799	48.04708	48.05181	48.05163	.00455
35	.59837	49.48964	49.49450	49.49433	.00469
. . .	. . .	. . .	. . .	. . .	. . .
46	.67196	65.35780	65.36382	65.36397	.00617
47	.67849	66.80036	66.80650	66.80666	.00630
48	.68690	68.24292	68.24919	68.24936	.00644
49	.50000	69.68549	69.69188	69.69206	.00657
50	.50000	71.12804	71.13458	71.13475	.00671
. . .	. . .	. . .	. . .	. . .	. . .

Table 2

Complete Ordering (Continued)

n	Optimal P_n	Optimal G_n	½-procedure $\tilde{G}_n$	Asymptotic $\frac{n}{\ln 2} - 1$	Difference $\frac{n}{\ln 2} - 1 - G_n$
60	.50000	85.55365	85.56177	85.56170	.00805
61	.50000	86.99621	87.00449	87.00440	.00819
62	.50000	88.43877	88.44721	88.44709	.00832
63	.50000	89.88133	89.88992	89.88978	.00845
64	.55377	91.32590	91.33264	91.33249	.00859
65	.57481	92.76646	92.77535	92.77518	.00872
. .	. . . .	. . . . .	. . . . .	. . . . .	. . . .
76	.61995	108.63462	108.64495	108.64482	.01020
77	.62904	110.07718	110.08762	110.08751	.01033
78	.63867	111.51974	111.53029	111.53021	.01047
79	.64672	112.96231	112.97297	112.97291	.01060
80	.65261	114.40487	114.41564	114.41561	.01074
81	.65673	115.84743	115.85831	114.85830	.01087
82	.65964	117.28999	117.30098	117.30100	.01101
83	.66176	118.73255	118.74365	118.74369	.01114
. .	. . . .	. . . . .	. . . . .	. . . . .	. . . .
90	.67282	128.83047	128.84238	128.84255	.01208

Multiplying (3.4) by e^{-2z}, we let

$$F(z) = e^{-z}y(z) = \sum_{n=0}^{\infty} \frac{A_n}{n} z^n$$

and obtain

$$(3.5) \qquad F(2z) = 2F(z) + 1 - 2(z + 1)e^{-z} \quad ,$$

$$(3.6) \qquad \sum_{n=0}^{\infty} \frac{A_n(2z)^n}{n!} - 2 \sum_{n=0}^{\infty} \frac{A_n z^n}{n!} = 1 + 2 \sum_{n=0}^{\infty} (-1)^n \frac{(n - 1)z^n}{n!} \quad .$$

Equating coefficients, we obtain for $n \geqq 2$

$$(3.7) \qquad A_n = \frac{(-1)^n(n - 1)}{2^{n-1} - 1} \quad ,$$

where A_0 and A_1 will be found later. Hence it follows that

$$(3.8) \qquad y(z) = \sum_{B=0}^{\infty} \frac{A_B}{B!} z^B \cdot \sum_{\alpha=0}^{\infty} \frac{z^\alpha}{\alpha!} = \sum_{n=0}^{\infty} \frac{z^n}{n!} \sum_{r=0}^{n} \binom{n}{r} A_r \quad .$$

Using (3.7) and the definition of $y(z)$, we obtain

$$(3.9) \qquad \tilde{G}_n = \sum_{r=0}^{n} \binom{n}{r} A_r \quad .$$

From (3.9) for $n = 0$ and 1 we find that $\tilde{G}_0 = A_0 = 1$ and

$$(3.10) \qquad \tilde{G}_1 = 1 = A_0 + A_1 = 1 + A_1 \Rightarrow A_1 = 0 \quad .$$

Hence we obtain the final result from (3.7), (3.9) and (3.10)

$$(3.11) \qquad \tilde{G}_n = 1 + \sum_{r=2}^{n} (-1)^r \binom{n}{r} \frac{r-1}{2^{r-1} - 1}$$

$$\sum_{r=2}^{n} (-1)^r \binom{n}{r} \frac{r-1}{1 - 2^{-r+1}} \; .$$

(In (3.11) we used the identity

$$\sum_{r=2}^{n} (-1)^r \binom{n}{r} (r-1) = 1.)$$

Asymptotic evaluation of $\tilde{G}_n$.

From (3.11) we also obtain for $\Delta\tilde{G}_n = \tilde{G}_n - \tilde{G}_{n-1}$ for $n > 2$

$$(3.12) \qquad \Delta\tilde{G}_n = \sum_{r=2}^{n} \frac{(-1)^r (r-1)}{1 - 2^{1-r}} \left[\binom{n}{r} - \binom{n-1}{r}\right] = \sum_{s=0}^{n-2} \frac{(-1)^s (s+1)}{1 - (\frac{1}{2})^{s+1}} \binom{n-1}{s+1}$$

$$= (n-1) \sum_{s=0}^{n-2} (-1)^s \binom{n-2}{s} \left[1 + (\tfrac{1}{2})^{s+1} + (\tfrac{1}{2})^{2s+2} + \ldots\right]$$

$$= (n-1)\left[\tfrac{1}{2}(\tfrac{1}{2})^{n-2} + \tfrac{1}{4}(\tfrac{3}{4})^{n-2} + \ldots\right]$$

$$= (n-1) \sum_{\alpha=1}^{\infty} \frac{1}{2^\alpha} \left(1 - \frac{1}{2^\alpha}\right)^{n-2} \; ;$$

if we set $\tilde{G}_1 = 1$ then (3.12) also holds for $n = 2$.

From (3.12), using the definition $\tilde{G}_1 = 1$, we obtain

(3.13) $$\tilde{G}_n - 1 = \sum_{\alpha=1}^{\infty} \frac{1}{2^{\alpha}} \sum_{i=2}^{n} (i - 1)(1 - \frac{1}{2^{\alpha}})^{i-2} \quad .$$

Letting $\theta = 1 - (1/2^{\alpha})$ and $j = i - 1$, we can write this for $n \geq 2$ as

(3.14) $$\tilde{G}_n = \sum_{\alpha=0}^{\infty} \frac{1}{2^{\alpha}} \sum_{j=0}^{n-1} j\theta^{j-1} = \sum_{\alpha=0}^{\infty} \left[\frac{1 - \theta^{n}}{1 - \theta} - n\theta^{n-1}\right] \quad .$$

We now use the Euler-MacLaurin sum formula for (3.14). For the analogous integral I, using x for α and letting $y = 1 - (1/2)^{x}$, so that $dx = dy/[(1 - y)\ell n2]$, we obtain

(3.15) $$I = \int_0^{\infty} \{2^{x}[1 - (1 - \frac{1}{2^{x}})^{n}] - n(1 - \frac{1}{2^{x}})^{n-1}\}dx$$

$$= \frac{1}{\ell n2} \int_0^1 \left[\frac{1 + y + y^2 + \dots + y^{n-1} - ny^{n-1}}{1 - y}\right] dy$$

$$= \frac{1}{\ell n2} \int_0^1 [1 + 2y + 3y^2 + \dots + (n - 1)y^{n-2}]dy = \frac{n - 1}{\ell n2} \quad .$$

The two correction terms for the Euler-MacLaurin sum formula yield 1/2 and 0; using the same analysis as in (3.15) it is easy to show that the remainder term is bounded by 1/2. Hence the asymptotic result for $\tilde{G}_n$ is

(3.16) $$\tilde{G}_n \approx \frac{1}{\ell n2} - 1 \approx 1.44269\, n - 1$$

where the adjusted constant is based on empirical results. (For $n = 50$ this gives 71.1345 and the exact result for $\tilde{G}_{50}$ is 71.1342, an error of less than 1/200 of 1%. The optimal result for $n = 50$ is 71.1280, an error in $\tilde{G}_n$ of about 1/100 of 1%.)

In this problem the best we can do is to point out that we need a minimum of at least $n - 1$ questions to separate all the n observations and (3.16) shows that for "1/2-procedure" on the average we need only about 44% more than this minimum.

It should be mentioned here that from a computational point of view the search for the optimal solution in our last setting represents a significant problem. The problem is that G_n as a function of n is so closely approximated by a linear function of n, and that for the fixed n, $G_n(p)$ is almost constant so that the search for p which yields the minimum to $G_n(p)$ is quite difficult. Thus, for $n \geqq 10$ the variation of p in the interval [.5,.7] does not change the first two decimals in $G_n(p)$. However, the difference $n(\ell n2)^{-1} - 1 - G_n$ (which show up in the 3^d decimal) tend to grow very slowly with n so that the correction term, namely the constant 1 in $G_n \doteq n(\ell n2)^{-1} - 1$, should actually be larger than one, say of the form $1 + \beta_n$ where β_n is a very slowly increasing function of n.

From Table 2 we empirically observe a cyclic pattern for the optimal p-value which ought to be described. The optimal p-value is always between .5 and a constant that appears to be close to $\ell n2 = .693...$.

For $2 \leqq n \leqq 5$ the optimal p is $.5$; for $6 \leqq n \leqq 9$ it increases; for $10 \leqq n \leqq 15$ it is again $.5$; for $16 \leqq n \leqq 23$ it increases; for $24 \leqq n \leqq 31$ it is again $.5$; for $32 \leqq n \leqq 48$ it increases; for $49 \leqq n \leqq 63$ it is again $.5$ and it increases for $n > 64$. For large r we conjecture that the optimal p will be $1/2$ for $3 \cdot 2^{r-2} \leqq n \leqq 2^r - 1$ and that it will increase between $1/2$ and some constant close to $\ell n2$ for $2^r \leqq n < 3 \cdot 2^{r-1}$ and that it will follow this type of cyclic pattern indefinitely. Furthermore, we conjecture that the small variation in the optimal G_n as p varies from $.5$ to $.7$ will persist for large values of n, so that the 1/2-procedure will always give an answer which is equal to the optimal G_n-value to 2 or 3 decimal places.

It is interesting to mention here that the natural generalization for the problem of complete ordering of the second optimality criterion from the selection problem (namely, maximizing the probability of a complete ordering in $n - 1$ steps) leads for $n \geqq 2$ to the equation

$$(3.17) \qquad P_n(p) = \sum_{k=1}^{n-1} \pi_{nk} P_k P_{n-k} = \sum_{k=2}^{n-2} \pi_{nk} P_k P_{n-k} + (\pi_{n1} + \pi_{n,n-1}) P_{n-1} \;,$$

where P_k denotes the probability of the complete ordering of a sample of size k in $k - 1$ steps and the π_{nk} are binomial probabilities:

$\pi_{nk} = \binom{n}{k}p^k(1 - p)^{n-k}$. For $n \leqq 6$ the optimal p (which yields the maximum for $P_n(p)$) is $1/2$; in the same way as above we can consider a "1/2-procedure" and denote by $\tilde{P}_n$ the corresponding probabilities and by P_n the maximal values over p of $P_n(p)$. It is easy to show that (3.17) implies for $n \geqq 2$ the inequalities

$$(3.18) \qquad \frac{1}{2}(\frac{2}{3})^{n-2} \leqq \tilde{P}_n \leqq P_n \leqq \frac{1}{2}(\frac{3}{4})^{n-2} \quad .$$

For the problem of the selection of the t largest with ordering out of n, we can easily write the equation for the expectation of the number of questions:

$$(3.19) \qquad G_n^{(t)}(p) = \frac{1 + \sum_{k=1}^{t} \pi_{nk}(G_k + G_{n-k}^{(t-k)}) + \sum_{k=t+1}^{n-1} \pi_{nk}G_k^{(t)}}{1 - p^n - q^n} \quad ,$$

where G_k denotes the minimal expectation for the problem of complete ordering, $G_k^{(t)} = \min_p G_n^{(t)}(p)$, and π_{nk} are binomial probabilities as before.

It is easy to see that

$$(3.20) \qquad G_t \leqq G_n^{(t)} \leqq G_t + \mu_n^{(t)} \quad ,$$

where the right-hand side corresponds to the expected total number of steps in the procedure "μ/G" in which we first select the t largest and then order them. Since G_t is of order t and $\mu_n^{(t)}$

is of the order $\sqrt{t}$ or less, with large values of t the optimal procedure for this problem of selection with ordering does not give the qualitative improvement over the procedure "μ/G" described above. However, our conjecture is that the optimal value of $p_n^{(t)}$ for the selection with ordering is between $t/2n$ and t/n and that there exists a constant ν such that $G_n^{(t)} - G_t \leqq \nu$ for all n and t.

Reference

Ross, S.M. [1970], Applied Probability Models with Optimization Applications, San Francisco: Holden-Day.

Reports of the Center for Research on Organizational Efficiency
(part of The Economic Series, IMSSS)

295. "On Partitioning of a Sample with Binary-Type Questions in Lieu of Collecting Observations," by Kenneth J. Arrow, Leon Pesotchinsky and Milton Sobel.